The Thesis Committee for Todd James Olson
Certifies that this is the approved version of the following thesis:

Evolution in one of the world's most abundant organisms, *Prochlorococcus*

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Mark Kirkpatrick

Lauren Meyers

Evolution in one of the world's most abundant organisms, *Prochlorococcus*

by

Todd James Olson, B.S.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

August 2016

Abstract

Evolution in one of the world's most abundant organisms, *Prochlorococcus*

Todd James Olson, M.A

The University of Texas at Austin, 2016

Supervisor: Mark Kirkpatrick

Prochlorococcus is one of the most abundant organisms on the planet, and this abundance has the potential to give it unique evolutionary properties. Very large idealized populations experience very little drift. This relationship has caused many authors to assume that *Prochlorococcus* genomes have little to no drift, but care needs to be taken to understand how real populations of *Prochlorococcus* differ from idealized populations and how this affects their evolution. Part of this thesis is devoted to understanding a parameter called the effective population size, which is meant illustrate the amount of drift in real, non-idealized populations. Using our knowledge of effective population size we can then investigate the relative affects of selection and drift throughout the genome. We can use knowledge about which sites evolve primarily by selection to better understand interpretations of common genome statistics. Finally, I apply my findings about interpretations of genome statistics to literature on the evolution of *Prochlorococcus* genome size.

Table of Contents

Evolution in one of the world's most abundant organisms, *Prochlorococcus*

INTRODUCTION

With an estimated global population of 10^{27} cells (Flombaum *et al*. 2013), *Prochlorococcus,* a globally distributed marine cyanobacterium, is one of the most abundant organisms on the planet. Due to its massive population size, *Prochlorococcus* may evolve unlike any other organism on earth. In addition to its effect on evolutionary dynamics, the extreme abundance of *Prochlorococcus* greatly affects the global ecosystem. *Prochlorococcus* fixes 4 gigatons of carbon a year, approximately the same amount as croplands globally (Flombaum *et al*. 2013, Biller *et al*. 2014, Huston & Wolverston 2009). Interestingly, *Prochlorococcus* has a highly reduced genome (as small as 1.66 Mb) (Kashtan 2014). Many other organisms with small genomes have extremely small effective population sizes and the cause of genome reduction in *Prochlorococcus* is still not well understood.

The efficacy of selection tends to increase with population size, such that alleles evolving primarily by drift in a small population may evolve primarily by selection in a larger population. For organisms with very large populations, there may be very few alleles evolving neutrally in their genome. A measure known as effective population size gives us more information about the strength of selection than census population size, and these values can differ significantly. In humans, for example, the census population size is around 7 billion, while the effective population size is under 10,000 (Tenesa *et al*. 2007). In *Prochlorococcus,* effective population size estimates differ by 7 orders of

magnitude, but all are many orders of magnitude lower than the census population size. In order to understand how *Prochlorococcus* evolves, it is important to recognize why measures of its effective population size differ from the census population size and where uncertainty in these measures comes from. The literature on *Prochlorococcus* genome evolution often fails to adequately address these questions.

In this paper I will review the literature on effective population size in *Prochlorococcus* evolution and discuss the implications of these results for understanding its evolution. In the first section I will discuss the concept of effective population size, with an emphasis on how to interpret common measures of this parameter. Next I will review the literature on calculating effective population size in *Prochlorococcus* and discuss potential problems with these calculations. In the next section, I investigate selection strength throughout the genome and discuss implications of these findings on genome analysis. In the final section I investigate how knowledge about effective population size and selection strength through the genome affects our understanding of genome reduction.

EFFECTIVE POPULATION SIZE

Effective population size (often abbreviated N_e) is an important parameter in evolution. There are a number of definitions of N_e, but the distinctions between these definitions is often not considered in evolution literature. In particular, there is the *variance effective size*, which tells us about the current influence of genetic drift in a population, and the *coalescent effective size*, which gives us information about the long-

term history of neutral alleles in the population. Equations derived from the coalescent N_e concept are often used to estimate effective size, but the estimated value is then interpreted using the first definition. The coalescent N_e and variance N_e often differ, so we need to take care in understanding where these differences come from. In this section I'll review the concept of two types of effective population size, variance N_e and coalescent N_e and explain how and why they can differ.

In an idealized evolutionary model (known as the Wright-Fisher model), selection acts more effectively on alleles evolving in larger populations than smaller ones (Wright 1931). This is because of the relationship between genetic drift and population size. Genetic drift is a process which causes allele frequencies to fluctuate due to random sampling. In a Wright-Fisher population, drift decreases with increasing population size. In other words, alleles experience less drift in larger populations and evolve more deterministically in accord with their selective benefit or detriment to the organism. Selection is less effective when there is more drift because randomness is playing a larger role in determining allele frequencies. In real, non-idealized populations the relationship between population size and drift is more complicated. The strength of genetic drift is affected by age structure, sex distribution, and spatial structure, among other factors (Charlesworth 2009). The same allele can evolve with different amounts of drift in two populations with the same census size if they differ in one or more of these factors. Since census population size by itself doesn't give us information about the efficacy of selection, Sewall Wright (1931) introduced the concept of effective population size.

One framework for defining effective size is that it is the number of individuals in a Wright-Fisher population that would exhibit some property seen in the population in question (Crow & Kimura 1970). Depending on the property of interest, one can define different types of effective population size. If we want to know the relative importance of present levels of drift in a real population we should be concerned with the *variance effective size* of that population. In this case the property to be compared between the real and idealized population is variance of an allele's frequency through time. Imagine going into a real population with census size N_c and finding that the variance of a particular allele's frequency in time is v. If one constructs an idealized population in which the variance of that same allele is equal to v, the size of the idealized population (N_e) would be the variance effective size of that real population. Due to the difficulty of measuring an allele's frequency through time, estimates of N_e are usually based on a different definition of the effective size.

In a more commonly used definition of N_e, the property in question is the coalescence time of alleles. Coalescent time is the expected time since randomly sampled alleles had a common ancestor. The effective population size statistic derived from this method is known as the *coalescent effective size*. This effective size is much easier to measure than the variance effective size, one only needs to know the mutation rate (μ) and neutral diversity (π) of your population in question. The coalescent effective size (N_e) is estimated using the equation, $N_e = \pi/2\mu$ (for diploids the equation is $N_e = \pi/4\mu$) (Lynch and Conery 2003). This equation comes from the fact that any differences in neutral alleles between individuals must have arisen due to mutations since their most recent

common ancestor (MRCA). The expected number of differences between two individuals is proportional to the time since coalescence (t_c) and the mutation rate ($\pi = 2\mu t_c$). The factor 2 comes about because mutations can happen along both alleles lineages. In an idealized Wright-Fisher population, the expected time to this MRCA is N, the population size. Therefore in an ideal population, $\pi = 2\mu N$ and $N=\pi/2\mu$. If you measure π and μ in a real population, you can calculate the number of individuals in an idealized population that would generate those data. This number is the coalescent effective size of your real population.

There are a number of different ways the variance effective size can differ from the coalescent effective size. This can sometimes cause confusion, as authors will often refer to both parameters as *the* effective population size. When conditions are met to cause them to differ, one can't assume that calculations of N_e from the coalescent reveal anything about current levels of genetic drift, which are determined by variance effective size. One way for these parameters to diverge is if population size is not constant. Coalescent effective size reflects the long-term history of alleles, and is sensitive to historical variation in census population size. Variance effective size, on the other hand, reflects only the current behavior of alleles and is insensitive to historical changes in census population size. In *Drosophila melanogaster*, for example, recent adaptations suggested selection was quite efficient and drift minimal, which would imply a large variance effective size. Calculations of effective size based on the neutral diversity, on the other hand, suggested a much smaller N_e (Karasov 2010). The reason for this discrepancy is that *Drosophila* had a relatively recent population bottleneck (Thornton

and Andolfatto 2006). The N_e calculated from neutral diversity reflects the long-term history of alleles, and was affected by the period in which *Drosophila* populations were relatively small. The variance N_e on the other hand, was determined only by the current, relatively large, population size. By only looking at the N_e calculated from neutral diversity, one would overestimate the influence of genetic drift on alleles in this population.

Historical changes in the mutation rate can also generate error in the calculation of N_e from neutral diversity. The derivation of $N_e = \pi/2\mu$ assumes that mutation rate remains constant. A mutation rate which was historically lower than measured will cause you to underestimate the time since coalescence, giving you an underestimate of N_e. Conversely, a historically higher mutation rate will cause you to overestimate N_e. Differences between mutation rates in lab cultures and wild populations will have a similar effect. If μ is measured in the lab, but is higher in wild populations, you will overestimate N_e and vice versa.

Finally, the identification of truly neutral sites presents a practical problem for calculating coalescent N_e. If the putatively neutral sites are under purifying selection, the estimated N_e will be lower than the true N_e. Purifying selection will cause a reduction in diversity at those sites. If these sites are assumed to be neutral, this will be interpreted as a shorter coalescence time, which would correspond to a smaller N_e. Conversely, if the chosen sites are under diversifying selection, the estimated N_e will be higher than the true N_e. All of these potential issues in calculating effective population size contribute to the disagreement about N_e in *Prochlorococcus*. In the next section I'll review the literature on

effective population size in *Prochlorococcus* and consider how some of these issues might affect estimates of this parameter.

N_E IN *PROCHLOROCOCCUS*

Using the coalescent method, *Prochlorococcus* has been estimated to have an effective population size between 10^6 to 10^{13} (Price and Arkin 2015, Kashtan *et al*. 2014). This is many orders of magnitude smaller than the census population size of 10^{27}. In addition, the estimates themselves vary over 7 orders of magnitude. In the next section I'll address two major questions about these calculations: what are possible sources of uncertainty in calculations of *Prochlorococcus* effective population size, and what causes the calculated N_e to be so much smaller than the census population size?

In order to calculate $N_{e,}$ Kashtan *et al*. used 3rd position codons as their putatively neutral sites. They then used a mutation rate of 10^{-10} to get their N_e of $1.5*10^9$. This was just for the e9312 ecotype, which was the focus of their sampling effort. Depending on the time of sampling, the e9312 ecotype is estimated to take up between 78% and 92% of the total abundance of *Prochlorococcus*, so the total species N_e may be slightly higher. Price and Arkin (2015) used average sequence difference at all sites within a subpopulation and a range of mutation rate estimates, from $5*10^{-10}$ to $2*10^{-9}$ to get their N_e estimate, which ranges from $5*10^6$ to 10^8. Uncertainty of these estimates is due to the previous issues mentioned: the constant population size assumption, the calculations reliance on mutation rate, and the difficulty of identifying neutral sites.

All estimates of N_e using the coalescent method rely on the assumption that population size has been constant since the time since two alleles shared a common ancestor, approximately N_e generations. Kashtan *et al*. (2014 SM) estimate that the most recent common ancestor of their focal ecotype was alive approximately 100 million years ago. Assuming ~200 generations a year, this is equal to $2*10^{10}$ generations since the MRCA. If these estimates are correct, *Prochlorococcus* must have an N_e less than $2*10^{10}$ in order for the coalescent method to accurately reflect their real N_e. If the true N_e is higher than $2*10^{10}$, the coalescent N_e will give an underestimate of their true effective size.

Uncertainty in the N_e can also come about due to uncertainty in the measure of mutation rate. For *Prochlorococcus* this is particularly problematic. Mutation rates are generally determined from experimental cultures grown on solid media, but *Prochlorococcus* is a notoriously difficult species to grow in this way. Methods for estimating mutation rate in liquid media, though not ideal (Kissling *et al*. 2013), provide a rough estimate for μ in *Prochlorococcus*. Osburne *et al*. (2011) exposed *Prochloroccus* to antibiotics and measured the frequency of resistant strains. Because resistance to these particular antibiotics is caused by a single base pair mutation in other organisms, they could use this frequency to estimate an upper bound on *Prochlorococcus* mutation rate. They estimated that μ was between 10^{-7} and 10^{-10}. Using the lower bound of their mutation rate gives a coalescent N_e orders of magnitude lower than Kashtan's estimate. Like population size, mutation rate also must be constant since the coalescence of sampled alleles in the population. *Prochlorococcus* have lost numerous DNA repair

genes, whose loss may have significant effects on mutation rate (Marais 2008). Sun and Blanchard (2014) suggest that this loss happened in early lineages of *Prochlorococcus* after diversification from their sister species *Synechococcus*. This is likely much before coalescence in the sampled individuals, so mutation rate may have been relatively constant in the recent past. If the loss of any genes affecting mutation rate did happen relatively recently, the coalescent method would give an underestimate of the variance N_e.

The coalescent method also relies on accurate identification of neutral sites in the genome. For clarity I want to begin by defining "neutral" or "evolves neutrally". Population genetics theory tells us that alleles with small selective effect in small populations will evolve as if they were neutral. This was formalized by Motoo Kimura (1968) as a comparison between the product N_e x s and 1, where s is a measure of the strength of selection at a particular allele. The selection coefficient s can range from 0 to 1. An allele with $s = 0$ has absolutely no fitness effect on the organism and one with $s = 1$ has the most extreme possible effect, for example lethality or infertility. If N_e x $s < 1$ the allele will evolve as if it were neutral. This means that an allele with some fitness effect can be "neutral" or "evolve neutrally" if it is in a small enough population. In other words drift, migration and mutation will determine the frequency of that allele in the population.

This presents an obvious problem. Even if one knows the exact selective effect of a particular allele, the effective population size is still needed to determine whether that site evolves neutrally. To calculate N_e you need to identify neutral sites, but to identify neutral sites, you need N_e. Kashtan *et al*. (2014) acknowledge this issue, and

choose to use third position codons as putatively neutral sites. Third position codons generally don't change the amino acid translated, and are often assumed to have no selective effect. In fact, changes in the third codon position can have a small selective effect, and they may evolve non-neutrally in organisms with relatively large N_e (Hershberg & Petrov 2008). Thus Kashtan *et al*. argue that their estimate of N_e is a lower bound on the real N_e. They estimate that the real N_e may be orders of magnitude higher than their lower bound and could be as high as 10^{13}, which would make *Prochlorococcus* the organism with the world's largest effective population size (Kashtan SM). Price and Arkin use diversity at all sites for their calculation. This is not justified in the paper and it is unclear why they did so. Unless every site in the genome evolves neutrally, their use of genome-wide diversity almost certainly leads to an underestimate of N_e.

Regardless of these issues, the effective population size is clearly much lower than the census population size of 10^{27} individuals. There are two reasons for this, geographic isolation and background selection. Despite the fact that they are widely distributed throughout the ocean, only relatively small subpopulations have the opportunity to interact on relevant time scales. Using an ocean mixing model, Kashtan *et al*. (2014) estimate that around 10^{13} *Prochlorococcus* cells are likely to interact on the order of a few generations, giving a rough estimate of the size of an evolutionary relevant subpopulation.

Further reduction in N_e is likely due to background selection, a consequence of *Prochlorococcus'* tightly linked genomes (Price and Arkin 2015). Background selection occurs due to linkage between deleterious mutations and the rest of the genome.

Beneficial mutations that arise on a genomic background with many deleterious mutations are unlikely to be selected for if the selective effect of the deleterious mutations outweighs the advantage conferred by the beneficial mutation. This can reduce the effective population size of an organism by orders of magnitude (Good *et al*. 2014). The variance of an allele's frequency is increased because it can be linked to a variety of different genetic backgrounds of varying selective effect. Thus, this linkage reduces the variance N_e of a population. Price and Arkin (2015) conducted simulations to determine the effect that background selection has on coalescent N_e as well. Because neutral sites can also be linked to deleterious mutations, and individuals with these deleterious mutations will be selected out of the population, background selection reduces neutral diversity. This reduction in neutral diversity causes a decrease in the coalescent N_e. They found that this reduction in coalescent N_e was in close agreement with the reduction in variance N_e. In other words, the reduction in calculated N_e by background selection was in close agreement with expectation that alleles with selective coefficients smaller than $1/N_e$ will evolve neutrally.

Based on theory from Good *et al*. (2014) and Gordo *et al*. (2008), Price and Arkin use estimates of the selection coefficients throughout the genome, the genome size, the census population size, and mutation rates to estimate that the effective population size of *Prochlorococcus*. They suggest that N_e is on the order of 10^7. They claim that the sample size of 96 individuals used by Kashtan *et al*. was too small, and that this is what leads to the discrepancy between the N_e estimates. Their calculation of 10^7 does not take into account recombination, which has the potential to break up linkage in the genome and

raise N_e. Recombination rates in *Prochlorococcus* aren't well known, but moderate rates could potentially increase N_e around 5-fold. This still isn't enough to bring their estimate up to Kashtan's calculated N_e of $1.5*10^9$.

In conclusion, there is no clear agreement on the effective population size of *Prochlorococcus*. Estimates of effective population size range from 10^6 to 10^{13}, but real values could be higher due to problems with assumptions made in these calculations.

SELECTION STRENGTH AND D_N/D_S

Many genome analysis tools used to study evolution require you to determine which alleles are evolving neutrally and which are evolving non-neutrally. As mentioned previously, whether or not an allele at a particular locus evolves neutrally depends on both the effective population size and the selective effect of that allele on the organism. Therefore an allele that evolves neutrally in a population with low N_e may evolve non-neutrally in a population with larger N_e. Assumptions about what types of sites evolve neutrally that are valid in many organisms may not hold in *Prochlorococcus*, due to its large effective population size. We have already seen the importance of identifying neutral sites for estimating effective population size. Another analysis tool that depends on the identification of neutral sites is a statistic known as the d_N/d_S ratio. I'll begin by explaining d_N/d_S and its standard interpretation, and will then go on to examine how our knowledge of selection throughout the *Prochlorococcus* genome affects our understanding of this statistic. While this section focuses on d_N/d_S, I will also briefly

discuss the implications of selection strength throughout the genome of *Prochlorococcus* on phylogeny reconstruction.

The d_N/d_S ratio compares the ratio of so-called "synonymous" to "non-synonymous" substitutions between lineages of organisms. Standard interpretations of d_N/d_S are grounded in the assumption that the non-synonymous substitutions evolve primarily by selection and the synonymous substitutions evolve neutrally. Non-synonymous substitutions are mutations that affect the sequence of amino acids being translated by a particular gene. Changing even a single amino acid in a protein can significantly alter its biophysical properties, which is expected to have a selective effect on the organism (Depristo *et al*. 2005). Charlesworth and Eyre-Walker (2006) conducted a study to estimate the number of neutral non-synonymous changes in *E. coli*, and found that ~97.2% of these substitutions were evolving non-neutrally. *E. coli* has an effective population size of $\sim 10^7$. If *Prochlorococcus* has an effective population size of around 10^7, near the lower bound of estimates, this would imply a similar percentage of non-synonymous substitutions are evolving neutrally. Of course, this depends on the assumption that *E. coli* have the same distribution of fitness effects as *Prochlorococcus* which is unlikely. Regardless, most non-synonymous substitutions are likely to evolve non-neutrally in *Prochlorococcus*.

Synonymous substitutions are mutations that change the DNA sequence without affecting the amino acid sequence translated from that DNA. Synonymous substitutions are often assumed to be neutral, since they don't affect the structure of the translated protein. The distinction between synonymous and non-synonymous substitutions is used

in other genome analysis tools as well. As mentioned previously, calculations of N_e based on nucleotide diversity require the identification of neutral sites. Kashtan *et al*. (2014) used diversity at third position sites in codons for their calculation of N_e because changes at most third position codon sites are synonymous.

Though differences at synonymous sites are often assumed to be neutral, we know that changes at these sites can evolve by selection due to a phenomenon known as codon bias. The code linking DNA to amino acids is degenerate, such that multiple triplet sequences (known as codons) can code for the same amino acid. Despite not changing the sequence of encoded amino acids, these types of substitutions do change which tRNA molecule is used in translation. There is a bias for usage of particular codons in many organisms (Quax *et al*. 2015) and this is known as codon bias (Sharp & Li 1986). Evidence suggests that this bias is adaptive (Berg & Kerland 1997, Andersson & Kurland 1990), though the exact mechanism is still not known. It is difficult to determine the exact selective effect driving codon bias but Price and Arkin (2015) suggest an $s \sim 5 \times 10^{-9}$ to 10^{-4}. This is right on the threshold of neutrality depending on our exact estimate of N_e. If N_e is on the higher end of estimates ($\sim 10^{11}$), these substitutions will certainly evolve by selection. On the lower end ($\sim 10^{6}$), they may or may not evolve as if they were neutral, depending on the exact value of the selection coefficient.

There may also be a selective effect of nucleotide content throughout the genome of *Prochlorococcus*. *Prochlorococcus* genomes have a strong nucleotide bias; up to 70% of their genomes are made up of A's and T's (Batut *et al*. 2014). It has been suggested that this bias is caused by adaptive evolutionary forces (Partensky and Garczarek 2010).

Nitrogen is a limiting resource for *Prochlorococcus* and AT molecules, as a pair, have one less nitrogen than GC pairs. This nucleotide bias is more pronounced in high light strains, which are more limited in nitrogen than low light strains (Gilbert and Fagan 2010). Selection for nucleotide content at an individual site is likely to be quite small, but selection for a mutation leading to AT bias throughout the entire genome could be strong enough to overcome drift. The strains of *Prochlorococcus* with the most drastic AT bias have lost numerous repair genes that correct accidental GC → AT mutations (Sun and Blanchard 2014). Since strains that lost this repair gene would have less nitrogen costly genomes, this gene loss may itself be adaptive. Patterns in d_N/d_S in *Prochlorococcus* support this hypothesis, but only under an interpretation of the statistic that takes into account the importance of selection on synonymous substitutions. Before covering this hypothesis more in-depth, I want to explain the standard interpretation of d_N/d_S and illustrate how we might change this interpretation due to selection on synonymous substitutions.

The d_N/d_S ratio is usually interpreted as a measure of selective pressure on protein-coding genes. The rate of synonymous substitution, d_S, is assumed to reflect the sum effects of all non-adaptive evolutionary forces (drift, migration, and mutation) on that particular coding region. If there are more non-synonymous differences between two lineages than synonymous ones, it is assumed that selection must be driving diversification between the lineages. Therefore a d_N/d_S ratio larger than 1 is interpreted as a signal of positive selection. A d_N/d_S ratio near 0 indicates that there have been very few substitutions affecting the amino acid sequence between the lineages in question relative

to the neutral expectation. This suggests that the protein or proteins being coded by this region of the genome are under purifying selection. This particular interpretation of d_N/d_S relies on the assumptions about what types of sites evolve neutrally, but these assumptions don't necessarily hold in *Prochlorococcus* due to the fact that synonymous sites may not be evolving neutrally.

What sort of information does d_N/d_S provide, given that both codon bias and nucleotide preference allow for the possibility that selection is acting on synonymous substitutions? Given our lack of understanding about the mechanisms underlying codon bias, it is difficult to predict exactly how we should interpret the d_N/d_S statistic in organisms where it is present. Selection for the same synonymous codons (purifying selection on codon usage) between lineages would reduce d_S, thus elevating d_N/d_S above what would be expected under neutral selection on synonymous substitutions. Selection for different synonymous codons (positive selection on codon usage) would elevate d_S, reducing d_N/d_S. The evolutionary mechanisms producing these patterns are difficult to disentangle, given that synonymous sites can no longer act as a neutral comparison to correct for the combined effects of drift, mutation, and migration.

Selection for nucleotide preference is likely to produce a more distinct pattern in d_N/d_S than codon bias. While AT pairs may have an intrinsic fitness benefit, these mutations can also have a detrimental effect on the organism due to other selective mechanisms. There is a selective balance between the harm caused by these mutations and the benefit of a less costly genome. Most non-synonymous substitutions are expected to be deleterious, and many of these deleterious mutations have relatively large selective

effects (Eyre-Walker & Keightley 2007). On average, non-synonymous GC → AT mutations will be more detrimental to the organism than synonymous ones. Therefore, strains with many synonymous GC → AT substitutions and few non-synonymous ones are more likely to have an overall selective benefit. Selection for an AT rich genome should lead to an enrichment of synonymous substitutions, indicated by a relatively low d_N/d_S compared to a lineage without this nucleotide bias.

Sun and Blanchard (2014) analyzed d_N/d_S along lineages before and after *Prochlorococcus* diversified from its sister species *Synechococcus*. After splitting from *Synechococcus*, *Prochlorococcus* diversified into two major clades, the high-light (often abbreviated HL) and low-light (LL) groups. HL strains tend to be more nutrient limited, have smaller genomes, and stronger bias towards AT. They found that there were a large number of synonymous substitutions relative to non-synonymous ones in the HL *Prochlorococcus* branch directly after diversification from the LL group. They interpreted the low d_N/d_S as a sign of strong purifying selection. In fact, these data fit perfectly with the hypothesis that a low d_N/d_S is an indicator of increased AT substitutions for a less nutrient costly genome, since this is the branch where *Prochlorococcus* adapted to nutrient poor waters. They also found that most of the DNA repair genes that fix GC → AT mutations were lost on this HL branch. This loss may have been concomitant with other adaptations that allowed *Prochlorococcus* to inhabit nutrient poor waters.

This analysis by Sun and Blanchard relies on their phylogeny, but selection at many sites in the genome could obscure the evolutionary history of their samples. We need to be careful in our interpretation of these results if we assume large N_e and strong

selection. Phylogenetic reconstruction requires some model of evolution. This model can be implicitly built in to the method, as is the case with parsimony. With maximum likelihood and Bayesian methods, an explicit model for evolution is chosen. In many cases, these models assume neutral evolution in the genes being analyzed. Sun and Blanchard used a maximum likelihood approach (TREE-PUZZLE), with a model in which similar sequences are assumed to have similar ancestry. The probability of substitutions is determined by mutational biases; there is no model for how selection may shape similarity and differences between lineages. If the sites being analyzed are under selection, similarity between sequences could also be due to convergent evolution. Castoe *et al*. (2009) showed that a phylogeny based on genes that underwent convergent evolution differs drastically from phylogenies reconstructed from other parts of the genome. If we assume that selection is the predominant evolutionary force throughout most of *Prochlorococcus* genome, similarity in sequences may reflect similar selection pressures rather than common ancestry. If this is the case, the initial HL branch may consist of many different lineages independently losing DNA repair mechanisms and adapting to nutrient poor water.

GENOME REDUCTION

Prochlorococcus has one of the smallest genomes of any bacteria (1.65-2.7Mb) (Partensky & Garczarek 2009). This initially puzzled scientists. Studies on in other bacteria with small genomes suggested that genetic drift was the cause of genome reduction, such that bacteria with large N_e should have relatively large genomes (Kuo,

Moran & Ochman 2009). Multiple hypotheses have been proposed to explain genome reduction in *Prochlorococcus*. These are roughly divided into adaptive explanations and non-adaptive explanations. Non-adaptive explanations include the mutation rate hypothesis and Muller's ratchet. Muller's ratchet is used to explain genome reduction in other bacteria, such as *Buchnera aphidicola* (Kuo, Moran & Ochman 2009). Muller's Ratchet occurs in populations with high levels of genetic drift so it is unlikely to explain genome reduction in *Prochlorococcus* (Batut *et al*. 2014). The mutation rate hypothesis is based on theory that suggests that an elevated mutation rate can lead to gene deletions, even when the deletions are detrimental to the organism's fitness (Marais *et al*. 2008). A problem with this hypothesis is that it doesn't provide an explanation for why mutation rate increased in the first place, and why it has remained high. This issue may be resolved by the observation that a heightened GC → AT mutation rate may be an adaptation to produce less costly genomes.

The adaptive explanations include the Black Queen and genome streamlining hypotheses. The Black Queen hypothesis states that genes can be discarded if the tasks those genes carried out were fulfilled by other organisms in the community (Morris, Lenski & Zinser 2012). This is also used as an explanation for why *Prochlorococcus* does so poorly in lab cultures. In order for the Black Queen hypothesis to fully explain genome reduction in *Prochlorococcus* it would have to be shown that all discarded genes carried out some function fulfilled by other organisms in their environment. This hypothesis also fails to explain why some strains of *Prochlorococcus* have smaller genomes than others. Comparative studies of the ecological communities in different

Prochlorococcus strains with varied gene losses may help elucidate the role of the Black Queen hypothesis in genome reduction.

One of the more popular explanations for *Prochlorococcus* reduced genome is the genome streamlining hypothesis. The genome streamlining hypothesis states that a small genome is selectively favored in nutrient-poor aquatic environments (Dufresne, Garczarek, and Partensky 2005, Rocap *et al*. 2003). In addition to the direct nutrient costs of producing DNA, a small genome allows for a smaller cell. A small cell is adaptive in nutrient poor water because the higher surface to volume ratio allows for more efficient nutrient uptake. Despite the advantages of a small genome, discarding genes may also detrimentally affect fitness. The genome streamlining hypothesis requires that the benefit of a smaller genome outweighs the cost of losing the function of discarded genes.

Sun and Blanchard (2014) attempted to answer this question by comparing reduced *Prochlorococcus* genomes with less reduced ones to determine which genes were discarded. Once they identified the discarded genes they calculated d_N/d_S ratios in these genes in branches where they were retained. The d_N/d_S ratio in these discarded genes was less than 1, but higher on average than non-discarded genes. Their interpretation of this result is that these genes were under relaxed purifying selection, and are less essential than other genes with lower d_N/d_S. This was used as support for the genome streamlining hypothesis, since we would expect that discarded genes would be less essential than retained ones. It is difficult to interpret these results given that we know about d_N/d_S in *Prochlorococcus*. A heightened d_N/d_S could still be a sign of relaxed purifying selection. It could also indicate a lower synonymous substitution rate, a sign of

constraint on codon choice. Clearly d_N/d_S is not a very useful statistic in this case. This confusion doesn't dismiss the Genome Streamlining hypothesis, but it does put some of its supporting evidence in question.

In an analysis of these 4 hypotheses: Muller's Ratchet, Black Queen, high mutation rate, and genome streamlining, Batut *et al*. (2014) suggest that genome streamlining is the strongest explanation of genome reduction in *Prochlorococcus*. On the other hand they argue that the loss of DNA repair genes presents a problem for this hypothesis. They claim that this gene loss doesn't fit with the expectation that discarded genes will in general be non-essential. In other words they believe that the benefit of reducing the genome by removing repair genes is less than the cost of a heightened mutation rate. As I've argued, this heightened GC → AT mutation rate may actually be a benefit in itself. Batut *et al*. acknowledge this possibility, but discard this idea because they claim that the loss of these repair genes doesn't correlate with AT content. They cite a study by Viklund *et al*. (2012) that found no correlation between AT content and presence of repair genes in alphaproteobacteria. In fact, this study only considered the genes mutT and agt. The highly reduced, high AT content *Prochlorococcus* strains also lost a gene called mutY, and the loss of both mutT and mutY causes a 1,000-fold increase in GC → AT mutations in *E. coli* (Horst *et al*. 1999). In addition, the lack of correlation between repair genes and actual AT content could be due to other selective forces. Bacteria with a biased mutation towards AT may still not have an AT-rich genome if these mutations are not selectively advantageous. In the case of *Prochlorococcus* there is a mutational bias towards AT *and* these mutations are retained due to selection.

Muller's Ratchet is unlikely to explain genome reduction in *Prochlorococcus* but the other three hypotheses, high mutation rate, Black Queen, and genome streamlining all have some support. These hypotheses are not mutually exclusive, and they could all explain genome reduction to varying degrees. The Black Queen hypothesis could be used to identify non-essential genes that can be easily discarded for a more streamlined genome. A heightened mutation rate could cause some genes to be discarded even if there is a net benefit to retain them. Our understanding of how to interpret d_N/d_S ratios in *Prochlorococcus* raises questions and answers others. In order to assess hypotheses of genome reduction we need to be careful to use or develop metrics that take into account its large effective population size and the effect this has on its evolution.

CONCLUSION

The peculiar biology of *Prochlorococcus* provides many interesting avenues of research for evolutionary biologists. Its massive population size immediately suggests highly effective selection, but research into its effective population size indicate that this might be overstated. On the other hand, many of these studies fail to take into account a number of subtleties that distinguish different conceptions of effective population size. Regardless, selection is quite strong in *Prochlorococcus* and this forces us to change our perspective on how we think about its evolution and what tools we use to study it. This change in perspective resolves some questions regarding the evolution of its reduced genome and raise new ones. *Prochlorococcus* is an extremely important player in the global ecosystem. Temperature is the best predictor of its abundance (Flombaum *et al*. 2013), suggesting that global climate change could affect the distribution of

Prochlorococcus throughout the ocean, which in turn would impact global carbon fixation. Understanding how it evolves is therefore not just a scientifically interesting question, but an environmentally important one as well.

References

Andersson, S. G., & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiological Reviews*, *54*(2), 198–210. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC372768/

Batut, B., Knibbe, C., Marais, G., & Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. *Nature Reviews Microbiology*, *12*(12), 841–850. http://doi.org/10.1038/nrmicro3331

Berg, O. G., & Kurland, C. G. (1997). Growth rate-optimised tRNA abundance and codon usage. *Journal of Molecular Biology*, *270*(4), 544–550. http://doi.org/10.1006/jmbi.1997.1142

Biller, S. J., Berube, P. M., Lindell, D., & Chisholm, S. W. (2014). Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology*, *13*(1), 13–27. http://doi.org/10.1038/nrmicro3378

Castoe, T. A., de Koning, A. P. J., Kim, H.-M., Gu, W., Noonan, B. P., Naylor, G., … Pollock, D. D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, *106*(22), 8986–8991. http://doi.org/10.1073/pnas.0900233106

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. http://doi.org/10.1038/nrg2526

Crow, J. F., & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. http://doi.org/10.2307/1529706

DePristo, M. a, Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews. Genetics*, *6*(9), 678–687. http://doi.org/10.1038/nrg1672

Dufresne, A., Garczarek, L., & Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology*, *6*(2), R14. http://doi.org/10.1186/gb-2005-6-2-r14

Eyre-Walker, a, & Keightley, P. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–8. http://doi.org/10.1038/nrg2146

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., … Martiny, A. C. (2013). Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(24), 9824–9. http://doi.org/10.1073/pnas.1307701110

Gilbert, J. D. J., & Fagan, W. F. (2011). Contrasting mechanisms of proteomic nitrogen thrift in Prochlorococcus. *Molecular Ecology*, *20*(1), 92–104. http://doi.org/10.1111/j.1365-294X.2010.04914.x

Good, B. H., Walczak, A. M., Neher, R. A., & Desai, M. M. (2014). Genetic Diversity in the Interference Selection Limit. *PLoS Genetics*, *10*(3). http://doi.org/10.1371/journal.pgen.1004222

Gordo, I., Navarro, A., & Charlesworth, B. (2002). Muller's ratchet and the pattern of variation at a neutral locus. *Genetics*, *161*(2), 835–848. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12072478

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annual Review of Genetics*, *42*, 287–299. http://doi.org/10.1146/annurev.genet.42.110807.091442

Horst, J. (1999). Escherichia coli mutator genes. *Trends in Microbiology*, *7*(1), 29–36. http://doi.org/10.1016/S0966-842X(98)01424-3

Huston, M. A., & Wolverton, S. (2009). The global distribution of net primary productivity. *Ecological Monographs*, *79*(3), 343–377. http://doi.org/10.1890/08-0588.1

Karasov, T., Messer, P. W., & Petrov, D. A. (2010). Evidence that adaptation in Drosophila is not limited by mutation at single sites. *PLoS Genetics*, *6*(6), 1–10. http://doi.org/10.1371/journal.pgen.1000924

Kashtan, N. *et al*. (2014), Single-Cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science (New York, NY)*, *344*(6182), 416–420. http://doi.org/10.1126/science.1248575

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), 624–626. http://doi.org/10.1038/217624a0

Kissling, G. E., Grogan, D. W., & Drake, J. W. (2013). Confounders of mutation-rate estimators: Selection and phenotypic lag in Thermus thermophilus. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, *749*(1-2), 16–20. http://doi.org/10.1016/j.mrfmmm.2013.07.006

Kuo, C., Moran, N. a, & Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity The consequences of genetic drift for bacterial genome complexity. *Genome Research*, (520), 1450–1454. http://doi.org/10.1101/gr.091785.109

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science (New York, N.Y.)*, *302*(5649), 1401–1404. http://doi.org/10.1126/science.1089370

Marais, G. A. B., Calteau, A., & Tenaillon, O. (2008). Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica*, *134*(2), 205–210. http://doi.org/10.1007/s10709-007-9226-6

Morris, J. J., Lenski, R. E., & Zinser, E. R. (2012). The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *mBio*, *3*(2). http://doi.org/10.1128/mBio.00036-12

Osburne, M. S., Holmbeck, B. M., Coe, A., & Chisholm, S. W. (2011). The spontaneous mutation frequencies of Prochlorococcus strains are commensurate with those of other bacteria. *Environmental Microbiology Reports*, *3*(6), 744–9. http://doi.org/10.1111/j.1758-2229.2011.00293.x

Partensky, F., & Garczarek, L. (2009). Prochlorococcus: Advantages and Limits of Minimalism. *Annual Review of Marine Science*, *2*(1), 305–331. http://doi.org/10.1146/annurev-marine-120308-081034

Price, M. N., & Arkin, A. P. (2015). Weakly deleterious mutations and low rates of recombination limit the impact of natural selection on bacterial genomes. *mBio*, *6*(6), 1–10. http://doi.org/10.1128/mBio.01302-15

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Molecular Cell*, *59*(2), 149–161. http://doi.org/10.1016/j.molcel.2015.05.035

Rocap, G., et. al (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, *424*(6952), 1042–1047. http://doi.org/10.1038/nature01947

Sharp, P. M., & Li, W.-H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol Journal of Molecular Evolution*, *24*(1-2), 28–38. http://doi.org/10.1007/bf02099948

Sun, Z., & Blanchard, J. L. (2014). Strong genome-wide selection early in the evolution of Prochlorococcus resulted in a reduced genome through the loss of a large number of small effect genes. *PLoS ONE*, *9*(3). http://doi.org/10.1371/journal.pone.0088837

Thornton, K., & Andolfatto, P. (2006). Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. *Genetics*, *172*(3), 1607–1619. http://doi.org/10.1534/genetics.105.048223

Viklund, J., Ettema, T. J. G., & Andersson, S. G. E. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Molecular Biology and Evolution*, *29*(2), 599–615. http://doi.org/10.1093/molbev/msr203